medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Rare variant effect estimation and polygenic risk prediction

Kisung Nam[1], Minjung Kho[1], Wei Zhou[2,3,4], Bhramar Mukherjee[5], Seunggeun Lee[1]

Abstract

Due to their low frequency, estimating the effect of rare variants is challenging. Here, we propose RareEffect, a method that first estimates gene or region-based heritability and then each variant effect size using an empirical Bayesian approach. Our method uses a variance component model, popular in rare variant tests, and is designed to provide two levels of effect sizes, gene/region-level and variant-level, which can provide better interpretation. To adjust for the case-control imbalance in phenotypes, our approach uses a fast implementation of the Firth bias correction. We demonstrate the accuracy and computational efficiency of our method through extensive simulations and the analysis of UK Biobank whole exome sequencing data for five continuous traits and five binary disease phenotypes. Additionally, we show that the effect sizes obtained from our model can be leveraged to improve the performance of polygenic scores.

[1] Graduate School of Data Science, Seoul National University, Seoul, South Korea
[2] Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA
[3] Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[4] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[5] Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Introduction

With the availability of extensive sequencing data in biobanks[1], the study of rare variants has become more feasible than ever before. Rare variants have been identified as potential causative factors for numerous complex diseases[2-8], and their exploration is crucial in unraveling the genetic risk factors of complex traits[9]. To identify the association between rare variants and complex traits, gene or region-based tests[10,11], including the Burden test, the sequence kernel association test (SKAT)[12] and its adaptive optimized version (SKAT-O)[13], have been commonly used. Recently, several methods, including STAAR[14], SAIGE-GENE[15], and SAIGE-GENE+[16] are developed to run region-based tests in biobank scale data.

To elucidate the effect of rare variants on complex diseases and traits and utilize them for risk prediction, in addition to calculating p-values for association tests, effect size estimation is required. However, estimating the effect size of rare variants remains a challenge. The low minor allele frequencies make single variant-based estimations unstable. Popular association tests like SKAT and SKAT-O are score tests, so only provide p-values. Although the Burden test approach provides a gene-burden effect size, this may not accurately reflect the true effect of rare variants in the presence of null variants and variants with opposing association directions.

To address the challenges, we introduce RareEffect, a novel method that estimates gene or region-based heritability and subsequently calculates each variant's effect size using an empirical Bayesian approach. Utilizing a variance component model, similar to that in the SKAT test, our method offers dual-level effect size estimation, region-level and variant-level, for enhanced interpretability. Unlike the Burden approach, our model flexibly estimates effect sizes in a data-driven manner. To reduce the computational burden for estimating variance components, we implemented the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM)[17], which leverages the low rank of the genetic relatedness matrix. We also utilize the strategy to collapse ultra-rare variants, as used in SAIGE-GENE+, to reduce the sparsity of the genotype matrix and improve power of estimating the effect of ultra-rare variants. For binary traits, we additionally apply a fast implementation of the Firth bias reduction method to stably estimate the effect sizes.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

From simulation studies, we showed that the proposed method is computationally fast and reliably estimates each gene heritability. We also showed that the proposed approach outperformed linear regression or ridge regression in terms of root mean squared error (RMSE) for estimating the individual-variant level effect sizes. From the UK Biobank (UKB) Whole Exome Sequencing (WES) data analysis of 5 quantitative traits and 5 binary disease traits, we demonstrate that exonic rare variants can explain substantial phenotypic variabilities, but the degree differs by phenotypes. We also showed that our approach has higher explanatory power in explaining the phenotype variability compared to models based on burden tests. These findings provide strong evidence for the practical utility of our method in leveraging rare variant data for risk prediction and heritability estimation.

Result

Overview of methods

Our proposed method encompasses three steps. The overview of the method is outlined in **Fig. 1** and is described in detail in the Methods section.

In step 1, we fit a null linear or logistic mixed-effect model without genotypes to estimate the covariate effects. This step involves fitting the model using the average information restricted maximum likelihood (AI-REML)[18] approach, which is utilized in the SAIGE[19] and GMMAT[20] framework. Residuals will be used as covariate-adjusted phenotypes in the subsequent steps.

In step 2, we model the effect of rare variants (MAF ≤0.01) as random effects and estimate the variance components, and hence heritability. To mitigate the computational demands, we adopt the Factored Spectrally Transformed Linear Mixed Models (FaST-LMM)[17]. Given $n \gg k$, FaST-LMM reduces the computation cost from $O(n^3)$ of conventional mixed model algorithm to $O(nk^2)$ where n is the number of samples in the dataset and k is the number of genetic variants in a single group (See **Methods**). Additionally, we leveraged the sparsity in genotypes. To incorporate the fact that genetic effects vary by functional annotations, we fit the model separately for distinct categories, and then combine them to calculate gene or region-level heritability. For whole exome analysis, we include three functional categories: (1) Loss-of-function (LoF) variants; (2) missense variants; and (3) synonymous variants. Within each

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

category, the ultra-rare variants, defined as those with a minor allele count (MAC) of lower than 10, are collapsed into a single variant, as employed in SAIGE-GENE+[16]. As estimating multiple variance components due to distinct functional groups is not feasible with FaST-LMM, we marginally calculate variance component for each functional group separately, and then adjust them using method of moments (MoM) approach (See **Methods**).

In step 3, following the estimation of variance components, we calculate the effect size of each variant using the Best Linear Unbiased Predictor (BLUP) estimates[21-24]. We further estimate the prediction error variance (PEV) of the effect size estimates to assess the reliability of the variant-level effect sizes. Additionally, using the estimated PEV, we can obtain confidence intervals for each variant-level effect size. For binary traits, we implemented Firth bias correction as a subsequent step. This correction mechanism mitigates bias and rectifies abnormal estimates, especially in scenarios where the case-control ratio is imbalanced. Recognizing the imperative of scalability in large-scale biobank data analyses, we developed a fast implementation of Firth correction, which reduces computation complexity from $O(Mnk_F)$ to $O(n_{nz}k_F)$ where M is the average number of iterations for convergence of Firth corrected beta, n_{nz} is the number of individuals with non-zero genotype, and k_F denotes the number of variants that needs to be corrected (See **Methods**).

UK Biobank WES data analysis

We applied our method to five quantitative traits (HDL cholesterol, LDL cholesterol, triglycerides, height, body mass index (BMI)) and five binary traits (breast cancer, prostate cancer, lymphoid leukemia, type 2 diabetes, and coronary atherosclerosis) in UKB. For LDL cholesterol (LDL-C) level, we adjusted the pre-medication levels by dividing the raw LDL-C level by 0.7 for individuals on cholesterol-lowering medication[25].

We computed gene-level effect sizes by leveraging the estimated heritability derived from the mixed effects model. Recognizing the inherent unsigned nature of heritability, we assign a sign by incorporating variant-level effect sizes of loss-of-function (LoF) variants within each gene. This allows us to discern the direction of the effect of the gene on the trait (See **Methods**). As expected, the gene-based association test p-values and the magnitude of the gene-level effect size showed a substantial correlation (**Fig. 2** and **Supplementary Figure 1**). Incorporating gene-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

level effect size and direction on top of the gene-based association tests can add significant value to genetic analyses. For example, the signed heritability clearly shows that the impairment of APOC3 function increases HDL cholesterol (HDL-C) level but decreases triglycerides (**Fig. 2(a)** and **2(c)**).

We estimated the variant-level effect sizes of exonic variants and presented two genes, *APOC3* and *SLC12A3,* on HDL-C levels as examples (**Fig. 3**). As expected, variants demonstrating significant associations, in terms of p-values, also exhibited larger effect sizes. Using RareEffect, we observed that the majority of variants in *APOC3* displayed positive effect sizes. The Burden and SKAT-O p-values from SAIGE-GENE+ were both highly significant (Burden p-value = 1×10^{-298} and SKAT-O p-value = 1×10^{-300}). In contrast, variants in *SLC12A3* exhibited both positive and negative effect sizes. Consequently, Burden p-value was not significant (Burden p-value = 0.01), but the SKAT-O p-value was significant (Burden p-value = 7×10^{-12}). This distinction in the directionality of effect sizes cannot be discerned through the burden approach.

We extended our analyses to estimate polygenic risk scores using rare variants with effect sizes estimated from RareEffect (PRS_{RE}). The UK Biobank data were randomly split into training and test sets (ratio = 8:2), and PRS_{RE} was constructed using the genes with p-values $< 2.5 \times 10^{-6}$ from SAIGE-GENE+ in the training set (See **Methods**). We included top 10 genes when the number of genes with p-values $< 2.5 \times 10^{-6}$ is smaller than 10. We further integrated these PRS_{RE} with PRS derived from common variants (PRS_{common}) to evaluate the practical utility of our approach. When combining the PRS from common and rare variants, we constructed a composite score, a linear combination of PRS from common and rare variants. We evaluated the predictive performance in terms of R^2. Our methods consistently exhibited superior prediction accuracy for all tested quantitative traits, compared to a comparative approach that relied on per-allele effect sizes derived from burden tests (**Fig. 4**, **Supplementary Table 1**, and **Supplementary Figure 2**).

The improvement became particularly pronounced when predicting lipid phenotypes among individuals deemed at high risk. For instance, when predicting HDL and LDL cholesterol levels, PRS_{RE} achieved R^2 of 0.4737 and 0.6287 when restricting individuals with top/bottom 0.5%

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

in terms of PRS_{RE}. When the composite score was used for risk prediction, R^2 were 0.6237 and 0.6645 when restricted top/bottom 0.5% individuals. In contrast, common variants only PRS model had lower R^2 (0.5417 for HDL and 0.5823 for LDL) for top/bottom 0.5% individuals. Notably, the sub-groups identified as high-risk by PRS_{common} and PRS_{RE} were substantially distinct (**Supplementary Figure 3**), underscoring the complementary nature of rare variants in detecting individuals at elevated disease risk. Additionally, we observed that our model showed higher predictability compared to the Burden approach which showed R^2 of 0.3683 and 0.2262 for HDL and LDL with top/bottom 0.5% in terms of PRS using burden score (PRS_{burden}), respectively.

Our method exhibited marginally lower predictive performance for binary traits, as measured by AUC, compared to the burden approach. We observed that for chosen binary traits, there were fewer genes associated with the trait, and their signals appeared weaker when contrasted with tested continuous traits. Nonetheless, our method offers potential benefits, as it can enhance predictability by combining PRS_{RE} with PRS_{common} and PRS_{burden}, which yielded better results. For instance, in the case of lymphoid leukemia, when evaluating individuals in the top/bottom 0.5% based on common PRS, the $PRS_{common} + PRS_{RE} + PRS_{burden}$ approach exhibited an AUC of 0.8649, whereas the $PRS_{common} + PRS_{burden}$ approach demonstrated an AUC of 0.8559, with the common-only approach yielding an AUC of 0.8559.

We further examined the relationship between phenotype outliers and the PRS in identifying individuals at high risk for common diseases[26]. We first defined the phenotype outliers as individuals with phenotype value exceeding a certain z-score cutoff and calculated the proportion of individuals with high PRS among phenotype outliers. Specifically, for LDL cholesterol levels, PRS_{RE} successfully pinpointed individuals at phenotypic extremes, who exhibited a tenfold higher likelihood of possessing a PRS_{RE} falling within 0.1st percentile compared to the baseline population (**Supplementary Figure 4**). PRS_{RE} and PRS_{common} utilize distinct set of variants and show minimal correlation (**Supplementary Table 2**). Therefore, integrating these models into a unified framework enables the identification of a significantly larger cohort at high risk than achievable through PRS_{common} alone.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Simulation study

To assess the predictive accuracy of our method, we conducted extensive simulations under diverse scenarios (see **Methods**) for both binary and quantitative traits. To mimic real data, we utilized actual genotypes from the UKB dataset, specifically the array-genotyped data for common variants (MAF≥0.01) and the UKB WES data for rare variants (MAF≤0.01).

We compared the performance of our method against other existing approaches such as linear regression, which is used for standard single-variant association test, or ridge regression in terms of RMSE. For quantitative traits, our method consistently demonstrated a lower RMSE of 0.1703 on average, outperforming the comparative methods which showed RMSE of 0.1770 (ridge regression) to 0.1881 (linear regression) (**Supplementary Table 3**) when estimating the effect size. For binary traits, our method also exhibited lower predictive error particularly in scenarios of low disease prevalence compared to ridge regression (**Supplementary Table 4**). Conversely, ridge regression showed marginally reduced error compared to our method in instances of high disease prevalence; however, the difference in accuracy remains modest.

Computation and memory cost

Analyzing 166,960 samples from the UKB WES data to estimate the effect size of the *DOCK6* gene, which contains 4,114 rare variants, we observed that the computation time for RareEffect with a simulated phenotype was approximately 90% lower than that of ridge regression. Specifically, RareEffect completed the analysis in 4.2 seconds, compared to 44.6 seconds for ridge regression (**Supplementary Figure 5).** The memory usage for RareEffect during the analysis was 1.14GB for *DOCK6* gene (**Supplementary Figure 6**). For binary traits, an additional step of performing Firth bias correction is required. We observed that the normal Firth bias correction process took 708 seconds to analyze a gene with 250 variants (after collapsing) across 342,409 individuals. However, by implementing our fast version of Firth correction, the computation time was dramatically reduced to 2.9 seconds (**Supplementary Figure 7**).

Discussion

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Our study introduces a novel method aimed at estimating the effect size of rare variants and can be extended to estimate gene-level effect size by employing a two-stage framework of generalized linear mixed models. By leveraging the variant-level effect size estimates obtained through our approach, we can examine the collective impact of rare variants within a gene and quantify their contribution to the overall heritability of complex traits.

In order to obtain accurate estimates for effect sizes while optimizing computational efficiency, we employed several techniques in our analysis. First, we utilized the FaST-LMM[17] algorithm to expedite computation and reduce memory usage. FaST-LMM leverages the spectral decomposition of the genetic relatedness matrix, allowing for efficient calculation of the variance component in mixed models. Second, we implemented the optimized version of Firth bias correction by utilizing the sparsity of genotype matrix and skipping the computation of hat matrix at every iteration. Third, we employed a collapsing strategy that reduces the sparsity of the data, similar to the approach employed in SAIGE-GENE+. These algorithmic approaches significantly accelerate the estimation process and enhance computational scalability, particularly when dealing with large-scale datasets and complex genetic analyses involving rare variants.

Beyond its immediate applications in effect size estimation, our proposed method offers significant potential for enhancing polygenic risk prediction models. Traditionally, these models have relied on common variants, often neglecting the valuable insights provided by rare variants. Our analysis reveals that the correlation between PRS_{RE} and phenotype values in the general population is not substantial (**Supplementary Figure 8**). However, we demonstrate that the RareEffect method effectively identifies individuals with high genetic risk. By incorporating our approach, we can substantially improve the predictive accuracy and precision of polygenic risk scores.

AI-based methods[27,28] have been developed to enhance the pathogenicity prediction of rare variants. Although these approaches help to identify effect sizes of pathogenic variants and can be used for risk prediction, they may not be as effective for identifying beneficial or gain-of-function variants. Additionally, the performance of these methods can be limited when applied to non-protein-altering variants. Our approach can accommodate the predictively

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

pathogenic variants identified by AI-based models by forming them into a separate category, thereby enhancing performance.

Our study, however, is not without limitations. While RareEffect demonstrates comparable or superior performance in estimating the effect size for simulated binary phenotypes, our evaluation revealed only marginal enhancements in predictive performance, as measured by the area under the curve (AUC), compared to the traditional common PRS across tested disease phenotypes. This could be attributed to the trait's reliance on a limited number of ultra-rare variants, which our method collapses into super-variants, thereby complicating the estimation of variant-level effect sizes for true causal variants. Despite its efficacy in effect size estimation, RareEffect's ability to improve prediction accuracy remains constrained, highlighting a potential area for future refinement and investigation for risk prediction of binary traits. Additionally, it is important to note that RareEffect is based on BLUP estimate, characterized by shrinkage properties, leading to biased estimates of effect sizes. However, despite this bias, RareEffect provides more stable estimates compared to unbiased methods like linear regression, which tend to be unstable for rare variant analysis due to the low allele frequency inherent in such variants.

In summary, our results demonstrate that incorporating information from rare variants enables the accurate estimation of gene-level and variant-level effect sizes, as well as the identification of high-risk individuals who might remain undetected by conventional polygenic risk scores (PRS) methods relying on common variants.

Method

Generalized linear mixed model

We denote the phenotype of the ith individual using y_i for both quantitative and binary traits in a study with sample size n. $\boldsymbol{X}$ represents the $n \times (p+1)$ vector with p covariates including the intercept and $\boldsymbol{G}_j$ is the $n \times k$ matrix representing the minor allele counts for k rare variants in gene or region j. The generalized linear mixed model (GLMM) can be expressed as:

$$g(\boldsymbol{\mu}) = \boldsymbol{X\alpha} + \boldsymbol{G}_j\boldsymbol{\beta}_j + \boldsymbol{b}$$

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

where $\boldsymbol{\mu}$ is the mean phenotype, $\boldsymbol{b} \sim MVN(0, \sigma_g^2 \boldsymbol{K})$ is the random effect, and $\boldsymbol{K}$ is an $n \times n$ genetic relatedness matrix (GRM). And g is the link function which is an identity function for quantitative traits with error term $\boldsymbol{\epsilon} \sim MVN(0, \sigma_e^2 \boldsymbol{I}_n)$ and a logit function for binary traits. The parameter $\boldsymbol{\alpha}$ is a $(p+1) \times 1$ vector of fixed effect coefficients and $\boldsymbol{\beta}_j$ is a $k \times 1$ vector of the random genetic effect.

Fitting the null generalized mixed model (step 1)

We used the average information restricted maximum likelihood (AI-REML) algorithm to fit the null GLMM (i.e., H_0: $\boldsymbol{\beta}_j = \boldsymbol{0}$) as in SAIGE step 1.

Estimation of the gene-level (region-level) heritability (step 2)

We estimate the effect size of rare variants using the following linear mixed model:

$$\tilde{\boldsymbol{y}} = \boldsymbol{G}_j \boldsymbol{\beta}_j + \boldsymbol{\epsilon}$$

where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^T = \boldsymbol{y} - \hat{\boldsymbol{y}}$ for quantitative traits, $\tilde{y}_i = \frac{d\eta_i}{d\mu_i}(y_i - \hat{\mu}_i) = \frac{1}{\hat{\mu}_i(1-\hat{\mu}_i)}(y_i - \hat{\mu}_i)$, a working residual from iteratively reweighted least squares (IRWLS) for binary traits, and $\hat{\mu}_i$ is the mean phenotype for individual i, which can be obtained from step 1. When obtaining $\boldsymbol{G}_j$, as in SAIGE-GENE+, we collapsed ultra-rare variants with minor allele count (MAC) ≤ 10 by each gene and functional group to reduce the sparsity. We further implemented an option to collapse all loss-of-function (LoF) rare variants into a single column, irrespective of their minor allele count, adopting the burden approach. This approach is predicated on the assumption that all rare LoF variants share a common effect size and direction. For binary traits, when we use working residuals, the variance of $\tilde{y}_i$ differs by individual, so we cannot apply the optimization technique of FaST-LMM. Therefore, we divide both sides by the square root of variance of $\tilde{y}_i$ to make the variance be the same across individuals. We estimate the effect size using the modified model, for binary traits:

$$\hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \tilde{\boldsymbol{y}} = \hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \boldsymbol{G}_j \boldsymbol{\beta}_j + \hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \boldsymbol{\epsilon}$$

where $\hat{\boldsymbol{\Sigma}}_\epsilon = \mathrm{diag}(\frac{1}{\hat{\mu}_i(1-\hat{\mu}_i)})$.

In this model, the prior distribution of $\boldsymbol{\beta}_j$ are assumed to follow $MVN(0, \tau\boldsymbol{\Sigma})$, and the noise $\boldsymbol{\epsilon}$ is assumed to follow $MVN(0, \psi \boldsymbol{I}_n)$ for quantitative traits, while we assume $\hat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \boldsymbol{\epsilon}$ follows

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

$MVN(0, \psi \boldsymbol{I}_n)$ for binary traits. When there is no prior knowledge of the correlation within $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ is set to be an identity matrix. But in general, $\boldsymbol{\Sigma}$ does not have to be an identity or a diagonal matrix.

To estimate the variance component parameters (τ, ψ), we use factored spectrally transformed linear mixed models (FaST-LMM) algorithm. Let $\widetilde{\boldsymbol{G}} = [\widetilde{\boldsymbol{G}}_1, \dots, \widetilde{\boldsymbol{G}}_k]$ be an $n \times k$ genotype matrix of the region with $\widetilde{\boldsymbol{G}}_j = \boldsymbol{G}_j$ for quantitative traits and $\widetilde{\boldsymbol{G}}_j = \widehat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \boldsymbol{G}_j$ for binary traits. The variance of $\widetilde{\boldsymbol{y}}$ (quantitative traits) or $\widehat{\boldsymbol{\Sigma}}_\epsilon^{-\frac{1}{2}} \widetilde{\boldsymbol{y}}$ (binary traits) can be written as

$$\boldsymbol{\tau} \widetilde{\boldsymbol{G}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{G}}^T + \psi \boldsymbol{I}_n$$

Traditional approaches to estimate the variance components require either calculating inverse matrix or conducting spectral decomposition of the $n \times n$ matrix $\widetilde{\boldsymbol{G}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{G}}^T$, so the time complexity is of $O(n^3)$. In contrast, FaST-LMM algorithm uses the fact that $\widetilde{\boldsymbol{G}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{G}}^T$ has rank at most k, so to reduce the computation complexity. Suppose $\boldsymbol{Z} = \widetilde{\boldsymbol{G}} \boldsymbol{L}$ is an $n \times k$ matrix where $\boldsymbol{L}$ is a Cholesky factor of $\boldsymbol{\Sigma}$. Then $\widetilde{\boldsymbol{G}} \boldsymbol{\Sigma} \widetilde{\boldsymbol{G}}^T = \boldsymbol{Z} \boldsymbol{Z}^T$. FaST-LMM carries out singular value decomposition on $\boldsymbol{Z}$ and calculate likelihood for (τ, ψ). With given $n \gg k$, calculation of $\boldsymbol{Z}$ and its singular value decomposition requires only $O(nk^2)$ of time complexity. And we further improved the computation efficiency utilizing the sparsity of $\boldsymbol{Z}$[29]. In biobank-scale data, n is in the hundreds of thousands, and k is in the tens to hundreds on average which means $n \gg k$.

Using the above optimization technique, we estimate the variance components by each group. Consider one group (LoF, missense or synonymous) in a single gene j in the model:

$$\widetilde{\boldsymbol{y}} = \boldsymbol{G}_{\cdot,j} \boldsymbol{\beta}_{\cdot,j} + \boldsymbol{\epsilon}$$

We first marginally estimate the maximum-likelihood estimator (MLE) of variance components $\tau_{LoF,j}$, $\tau_{mis,j}$, and $\tau_{syn,j}$. As the marginal estimates do not account for LD among variants in different groups, we adjust the estimates using method of moments (MoM) approaches[30]. The MoM estimator can be obtained by solving the following linear system:

$$\begin{bmatrix} \boldsymbol{T} & \boldsymbol{b} \\ \boldsymbol{b}^T & n \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \hat{\psi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{c} \\ \widetilde{\boldsymbol{y}}^T \widetilde{\boldsymbol{y}} \end{bmatrix}$$

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

where $\boldsymbol{T}$ is a 3×3 matrix for joint estimation or a scalar (1×1 matrix) for marginal estimation with entries $\boldsymbol{T}_{k,l} = \mathrm{tr}(\widetilde{\boldsymbol{G}}_k \boldsymbol{\Sigma}_k \widetilde{\boldsymbol{G}}_k^T \widetilde{\boldsymbol{G}}_l \boldsymbol{\Sigma}_l \widetilde{\boldsymbol{G}}_l^T)$ where $k, l \in \{1, 2, 3\}$ for joint estimation. $\boldsymbol{b}$ is a 3-vector for joint estimation or a scalar for marginal estimation with entries $b_k = \mathrm{tr}(\widetilde{\boldsymbol{G}}_k \boldsymbol{\Sigma}_k \widetilde{\boldsymbol{G}}_k^T)$, $\boldsymbol{c}$ is a 3-vector for joint estimation or a scalar for marginal estimation with entries $c_k = \widetilde{\boldsymbol{y}}^T \widetilde{\boldsymbol{G}}_k \boldsymbol{\Sigma}_k \widetilde{\boldsymbol{G}}_k^T \widetilde{\boldsymbol{y}}$, and $\hat{\boldsymbol{\tau}} = \begin{bmatrix} \widehat{\tau_1} \\ \widehat{\tau_2} \\ \widehat{\tau_3} \end{bmatrix}$ where (τ_1, τ_2, τ_3) denotes $(\tau_{LoF,j}, \tau_{mis,j}, \tau_{syn,j})$ for a single gene j, respectively.

After estimating the marginal and joint MoM estimator of variance components, we adjust the MLE of the marginal variance components by:

$$\hat{\tau}_{i,j} = \hat{\tau}_{MLE,mar,i,j} \times \frac{\hat{\tau}_{MoM,joint,i,j}}{\hat{\tau}_{MoM,mar,i,j}}$$

where $i \in \{LoF, mis, syn\}$. The MoM estimator of the variance component was not directly utilized in our study since the MoM estimator could yield negative variance components. In cases of the variance component estimated by the MoM is negative, we used marginal variance component without adjustment. Both MoM and likelihood-based approaches exhibited the similar trends of variance components (**Supplementary Figure 9**). Additionally, we assume that the variance explained by rare variants in a single gene is negligibly small compared to the total variance of $\tilde{y}$, therefore, $\hat{\psi} \approx Var(\tilde{y})$.

We estimated the heritability from rare variants of gene j using these adjusted variance components. In a joint model,

$$Var(\widetilde{\boldsymbol{y}}) = \boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T + \psi \boldsymbol{I}_n$$

where $\boldsymbol{G}_j = [\boldsymbol{G}_{LoF,j} \quad \boldsymbol{G}_{mis,j} \quad \boldsymbol{G}_{syn,j}]$ and $\boldsymbol{\Sigma}_j = \begin{bmatrix} \hat{\tau}_{LoF,j} \boldsymbol{\Sigma}_{LoF,j} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \hat{\tau}_{mis,j} \boldsymbol{\Sigma}_{mis,j} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \hat{\tau}_{syn,j} \boldsymbol{\Sigma}_{syn,j} \end{bmatrix}$.

Therefore,

$$\sum_{i=1}^{n} Var(\widetilde{y}_i) = \mathrm{tr}(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T) + n\psi \quad \text{(for quantitative traits)}$$

$$\sum_{i=1}^{n} Var(\widetilde{y}_i) = \mathrm{tr}(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T) + n\,\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_\epsilon) \quad \text{(for binary traits)}$$

Subsequently, the heritability from rare variants of gene j can be denoted as:

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

$$h_j^2 = \frac{\mathrm{tr}\left(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T\right)}{\mathrm{tr}\left(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T\right) + n\psi} \quad \text{(for quantitative traits)}$$

$$h_j^2 = \frac{\mathrm{tr}\left(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T\right)}{\mathrm{tr}\left(\boldsymbol{G}_j \boldsymbol{\Sigma}_j \boldsymbol{G}_j^T\right) + \psi\, \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}_\epsilon)} \quad \text{(for binary traits)}$$

Additionally, to determine the direction of the gene-level effect, we obtained the sign of the linear combination of the effect sizes of loss-of-function variants in a gene, weighted by their MAFs:

$$\text{(signed heritability of gene } j) = h_j^2 \times \mathrm{sgn}\left(\textstyle\sum_{j \in LoF} \beta_j\, MAF_j\right)$$

This measure gives the information of the magnitude of genetic effects from rare variants in a single gene and its direction of effects.

Estimation of the variant-level effect size (step 3)

The effect sizes at variant-level resolution are estimated using the adjusted variance components in the previous step by:

$$\widehat{\boldsymbol{\beta}}_j = \left(\widetilde{\boldsymbol{G}}_j^T \widetilde{\boldsymbol{G}}_j + \hat{\psi} \boldsymbol{\Sigma}_j^{-1}\right)^{-1} \widetilde{\boldsymbol{G}}_j^T \widetilde{\boldsymbol{y}}$$

for each gene or region.

We further estimated the prediction error variance (PEV) by:

$$PEV\left(\widehat{\boldsymbol{\beta}}_j\right) = \left(\widetilde{\boldsymbol{G}}_j^T \widetilde{\boldsymbol{G}}_j + \frac{\hat{\psi}}{\hat{\tau}} \boldsymbol{I}\right)^{-1}$$

for each gene or region. Using this PEV, we can obtain confidence intervals for effect sizes.

For binary traits, the Firth bias correction[31] is a more accurate method to estimate SNP effect sizes[32-34], particularly in cases marked by a significant case-control imbalance. We incorporate this correction into our analytical framework. Additionally, we introduce an L2 penalty term to account for the prior distribution of $\boldsymbol{\beta}$. The Firth corrected effect estimates can be calculated numerically by optimizing the following objective function:

$$\hat{\beta}_* = \mathrm{argmax}_\beta \left[\log L(\beta) + \frac{1}{2} \log |I(\beta)| - \frac{1}{2\hat{\tau}} \|\beta\|_2^2\right]$$

where L denotes the likelihood function and I is the Fisher information.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

To improve computational efficiency, we developed the fast implementation of Firth bias correction. First, we compute the hat matrix, $\boldsymbol{H} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{\frac{1}{2}}$, $\boldsymbol{W} = \mathrm{diag}(\hat{\mu}_i(1 - \hat{\mu}_i))$, only once rather than recalculating it at each iteration, assuming that the iterative weights ($\boldsymbol{W}$) change slowly as a function of the mean μ_i[20,35]. We compared, using both simulation and real data, the difference in effect size estimates between computing the hat matrix once and computing it at each iteration. Our findings indicate minimal differences in the effect size estimates (**Supplementary Figure 10**). Given that Firth correction is applied to each variant, we iterate the hat matrix calculation for k_F times, where k_F represents the number of variants that need to be corrected. This strategy results in a reduction in the computational complexity with Firth correction from $O(Mnk_F)$ to $O(nk_F)$, where M is the average number of iterations needed for Firth corrected effect size convergence. Second, we extend Firth correction to accommodate sparse genotype matrices. Specifically, we restrict our computations to individuals with non-zero genotypes when determining the score and Fisher information. This further reduces the time complexity of the Firth correction, now at $O(n_{nz}k_F)$, where n_{nz} denotes the number of individuals with non-zero genotype. Considering that we are estimating the effect size of rare variants with MAF < 0.01, we observed that $n_{nz} < 0.01n$ which implies that leveraging sparsity makes the estimation of effect sizes more than 100 times faster compared to a non-sparse approach.

We note that Firth correction is used when the absolute value of estimated effect size ($\hat{\beta}$) surpasses a predefined threshold. In this study, we used threshold of $\log 2 \approx 0.693$ for simulation studies and real data analysis.

Rare-variant PRS calculation

We further calculated polygenic risk scores using the variant-level effect sizes of rare variants (PRS_{RE}) in genes with gene-level p-value from SAIGE-GENE+ lower than 2.5×10^{-6}. PRS_{RE} of individual i can be calculated as:

$$PRS_{RE,i} = \sum_{j \in J} \boldsymbol{G}_{ij}\hat{\boldsymbol{\beta}}_j$$

where J denotes a set of genes with gene-level p-value lower than 2.5×10^{-6}. Additionally, we combined these PRS_{RE} with PRS_{common} only. We applied PRS-CS[36] to obtain the variant-

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

level weights for the calculation the PRS_{common}. We compared the predictive performance of the PRS_{RE} to PRS_{burden}. PRS_{burden} were obtained in two different ways: (1) collapsing all rare variants into one super-variant regardless of the functionality of the variants and (2) collapsing rare variants by functionality of the variants (LoF, missense, and synonymous). After collapsing, we fitted the following linear model to estimate the per-allele effect sizes:

$$\tilde{y} = \boldsymbol{G}_{burden,j}\boldsymbol{\beta}_{burden,j}$$

The burden PRS are also calculated as a linear combination of per-allele effect sizes and the collapsed genotypes of each individual.

We compared the predictive performance in terms of R^2 for quantitative traits, and the area under receiver operating characteristic curve (AUROC) for binary traits of the following linear models:

$$\tilde{y} \sim PRS_{composite}$$

$$\tilde{y} \sim PRS_{common} + PRS_{RE}$$

$$\tilde{y} \sim PRS_{common} + PRS_{burden}$$

where $PRS_{composite}$ is a linear combination of PRS_{common} and PRS_{RE} with weights trained from the training set.

UK Biobank data analysis

In this study, we used WES data of 393,247 White British participants in the UK Biobank. The UK Biobank is a UK-based prospective cohort of ~500,000 individuals aged 40 to 69 at enrollment. We split the train and test data 8:2 randomly for the PRS evaluation. We applied quality control (QC) procedures prior to the analysis. We first removed redundant samples and individuals with sex mismatch or sex chromosome aneuploidy. Additionally, we further removed variants with a missingness rate across individuals > 0.1, HWE p-value < 10^{-15}, and monomorphic variants. We generated group files, which define the list of variants in genes and its functional annotation, by using the loss-of-function transcript effect estimator (LOFTEE)[37]. We regarded a variant as loss-of-function (LoF) only in case of it is labeled as a high-confidence (HC) LoF variant, and variants with low-confidence (LC) were regarded as missense variants.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Using the data after QC, we applied our method to five quantitative traits (HDL cholesterol, LDL cholesterol, triglycerides, height, and body mass index) and five binary traits (breast cancer, prostate cancer, lymphoid leukemia, type 2 diabetes, and coronary atherosclerosis). We defined the disease by mapping ICD-10 codes to Phecodes using the PheWAS R package[38].

Simulation study

To generate outcome phenotypes, we used the following model for quantitative and binary traits:

$$y_i = X_{i1} + X_{i2} + G_{i,common}\beta_{common} + G_{i,rare}\beta_{rare} + \epsilon$$

$$logit\big(P(Y_i = 1)\big) = \alpha + X_{i1} + X_{i2} + G_{i,common}\beta_{common} + G_{i,rare}\beta_{rare}$$

where X_{i1} and X_{i2} are covariates, and $G_{i,common}$ and $G_{i,rare}$ are genotype vectors of common variants and rare variants of ith individual, respectively. The intercept α for binary traits is determined by the disease prevalence. The covariates X_{i1} and X_{i2} were simulated from Bernoulli(0.5) and $N(0,1)$, respectively. For common variant effect, we randomly selected L =30,000 LD-pruned common variants with MAF > 1%, and assumed that the effect size of single common variant follows $N(0,\frac{1}{L})$. We selected 10 causal genes in UKB WES 200k data for generation of phenotypes. We used eight different scenarios regarding rare variants: (1) proportion of causal variants, (2) effect size of causal variants, and (3) direction of effect within a single gene. For the proportion of causal variants, we assumed (1) 20% of LoF, 10% of missense, and 2% of synonymous variants, or (2) 30% of LoF, 10% of missense, and 2% of synonymous variants among rare variants that are not ultra-rare are causal. For ultra-rare variants, we assumed that the proportion of causal variants are three times higher than the non-ultra-rare variants, that is, (1) 60% of LoF, 30% of missense, and 6% of synonymous ultra-rare variants, or (2) 90% of LoF, 30% of missense, and 6% of synonymous ultra-rare variants are causal. Regarding the effect size of causal variants, we assumed that the absolute effect sizes of causal variants are (1) $|0.5\log_{10} MAF|$ for LoF variants, and $|0.25\log_{10} MAF|$ for missense and synonymous variants, or (2) $|0.3\log_{10} MAF|$ for LoF variants, and $|0.15\log_{10} MAF|$ for missense and synonymous variants. We further assumed that the effect directions are (1) the same among all causal variants in a single gene, or (2) same for 100% of LoF, 80% of missense, and 50% of synonymous variants, while remaining variants have the

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

opposite direction of effect. For eight combinations of scenarios, we repeated the simulation for 100 times.

Computation cost evaluation

We evaluated the computation time and memory usage using simulated data as described above, comprising 166,960 individuals of White British ancestry from the UKB WES 200k dataset. Additionally, we examined computation time and memory usage across subsets with sample sizes of 10k, 30k, 50k, and 100k. For each generative scenario, we reported the mean of 5 attempts for computation times and memory usage, comparing them with multiple linear regression, simple linear regression (as in GWAS), and ridge regression. The evaluation for linear regression and ridge regression was done using lm and glm functions in R, respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Data availability

The analysis results for 5 quantitative and 5 binary phenotypes of UKB WES data analysis results are available at: https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_effect_size.zip (variant-level effect size) and https://storage.googleapis.com/leelabsg/RareEffect/RareEffect_h2.zip (gene-level signed heritability).

Code availability

RareEffect is implemented as a part of SAIGE software, which is an open-source R package, available at https://github.com/saigegit/SAIGE. RareEffect is available in SAIGE version 1.3.7 or higher.

Author Contribution

K.N. and S.L. designed experiments. K.N. performed experiments and analyzed the UKB WES data. K.N. and S.L. implemented the software with input from W.Z.. M.K. and B.M. provided helpful advice. K.N. and S.L. wrote the manuscript with input from all co-authors.

Acknowledgements

This research was supported by the Brain Pool Plus (BP+) Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (2020H1D3A2A03100666) and the grants funded by the Ministry of Food and Drug Safety, Republic of Korea (Grant Number: 23212MFDS202). This research was conducted using the UK Biobank Resource under application number 45227.

References

1. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
2. Gibson, G. Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135-45 (2012).
3. Kryukov, G.V., Shpunt, A., Stamatoyannopoulos, J.A. & Sunyaev, S.R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proc Natl Acad Sci U S A* **106**, 3871-6 (2009).
4. MacArthur, D.G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823-8 (2012).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

5. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).
6. Rivas, M.A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet* **43**, 1066-73 (2011).
7. Saint Pierre, A. & Génin, E. How important are rare variants in common disease? *Briefings in Functional Genomics* **13**, 353-361 (2014).
8. Perrone, F., Cacace, R., van der Zee, J. & Van Broeckhoven, C. Emerging genetic complexity and rare genetic variants in neurodegenerative brain diseases. *Genome Med* **13**, 59 (2021).
9. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
10. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* **95**, 5-23 (2014).
11. Chen, W., Coombes, B.J. & Larson, N.B. Recent advances and challenges of rare variant association analysis in the biobank sequencing era. *Front Genet* **13**, 1014947 (2022).
12. Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* **89**, 82-93 (2011).
13. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* **91**, 224-37 (2012).
14. Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat Genet* **52**, 969-983 (2020).
15. Zhou, W. *et al.* Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. *Nat Genet* **52**, 634-639 (2020).
16. Zhou, W. *et al.* SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests. *Nat Genet* **54**, 1466-1469 (2022).
17. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* **8**, 833-5 (2011).
18. Gilmour, A.R., Thompson, R. & Cullis, B.R. Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440-1450 (1995).
19. Zhou, W. *et al.* Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nat Genet* **50**, 1335-1341 (2018).
20. Chen, H. *et al.* Control for Population Structure and Relatedness for Binary Traits in Genetic Association Studies via Logistic Mixed Models. *Am J Hum Genet* **98**, 653-66 (2016).
21. Robinson, G.K. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science* **6**, 15-32, 18 (1991).
22. Henderson, C.R. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics* **31**, 423-447 (1975).
23. Maier, R.M. *et al.* Improving genetic prediction by leveraging genetic correlations among human diseases and traits. *Nat Commun* **9**, 989 (2018).
24. Chen, C.Y., Han, J., Hunter, D.J., Kraft, P. & Price, A.L. Explicit Modeling of Ancestry Improves Polygenic Risk Scores and BLUP Prediction. *Genet Epidemiol* **39**, 427-38 (2015).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

25. Graham, S.E. *et al.* The power of genetic diversity in genome-wide association studies of lipids. *Nature* **600**, 675-679 (2021).
26. Fiziev, P.P. *et al.* Rare penetrant mutations confer severe risk of common diseases. *Science* **380**, eabo1131 (2023).
27. Sundaram, L. *et al.* Predicting the clinical impact of human mutation with deep neural networks. *Nat Genet* **50**, 1161-1170 (2018).
28. Gao, H. *et al.* The landscape of tolerated genetic variation in humans and primates. *Science* **380**, eabn8153 (2023).
29. Berry, M.W. Large-Scale Sparse Singular Value Computations. *International Journal of Supercomputer Applications and High Performance Computing* **6**, 13-49 (1992).
30. Pazokitoroudi, A. *et al.* Efficient variance components analysis across millions of genomes. *Nat Commun* **11**, 4020 (2020).
31. Firth, D. Bias Reduction of Maximum Likelihood Estimates. *Biometrika* **80**, 27-38 (1993).
32. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and binary traits. *Nat Genet* **53**, 1097-1103 (2021).
33. Wang, X. Firth logistic regression for rare variant association tests. *Front Genet* **5**, 187 (2014).
34. Dey, R., Schmidt, E.M., Abecasis, G.R. & Lee, S. A Fast and Accurate Algorithm to Test for Binary Phenotypes and Its Application to PheWAS. *Am J Hum Genet* **101**, 37-49 (2017).
35. Breslow, N.E. & Clayton, D.G. Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association* **88**, 9-25 (1993).
36. Ge, T., Chen, C.Y., Ni, Y., Feng, Y.A. & Smoller, J.W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat Commun* **10**, 1776 (2019).
37. Karczewski, K.J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434-443 (2020).
38. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Figures Legends

Fig. 1. Overview of RareEffect.

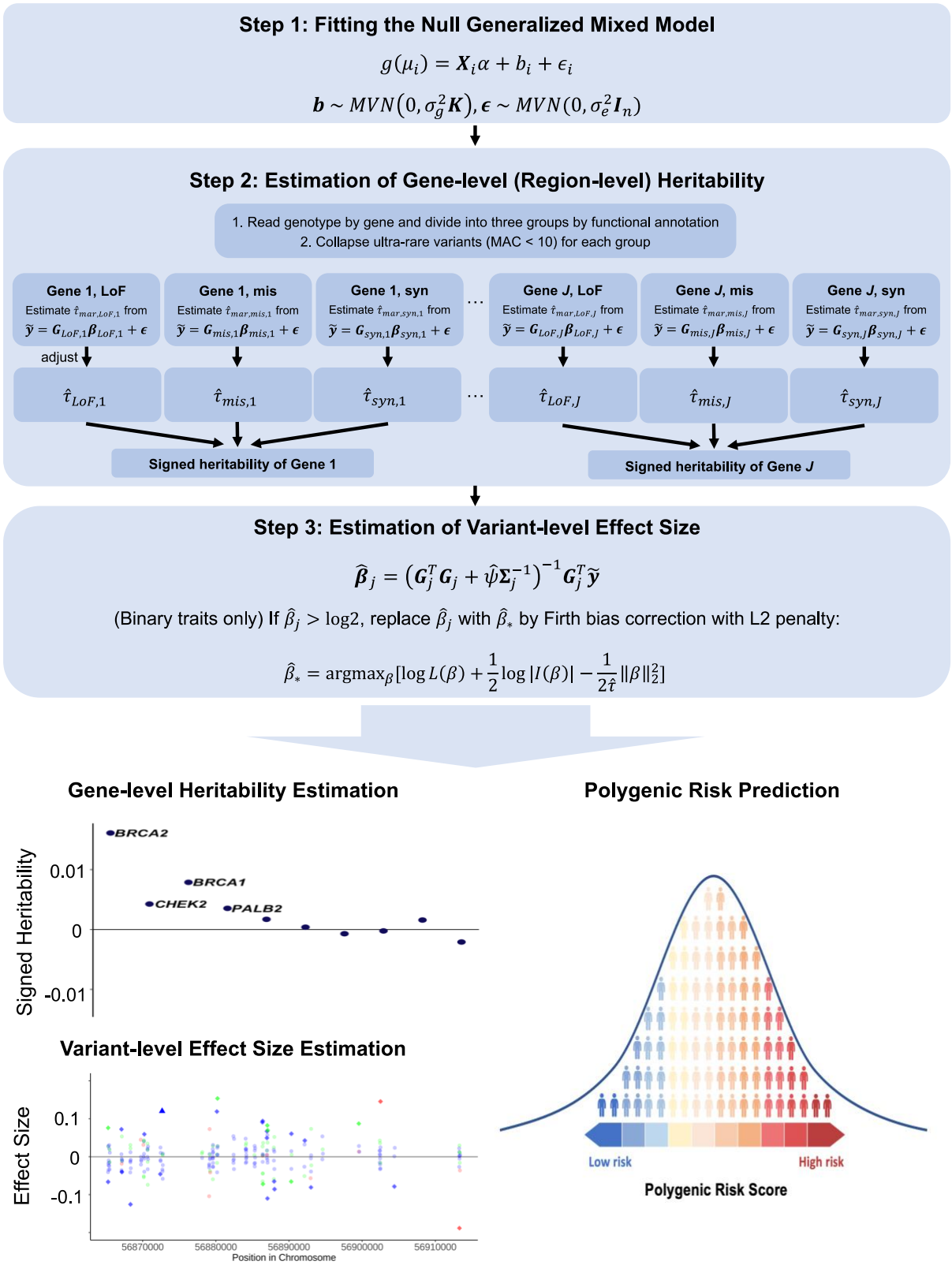
Fig. 2. Estimated signed heritability for lipid phenotypes using 392,748 White British samples in UK Biobank whole exome sequencing data

Fig. 3. Variant-level effect size on HDL cholesterol for variants in *APOC3* gene (chromosome 11) and *SLC12A3* gene (chromosome 16)

Fig. 4. Comparison of performance of risk prediction models for lipid phenotypes using 314,198 White British samples (80% of the whole White British samples) in UK Biobank whole exome sequencing data

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Fig. 1. Overview of RareEffect.



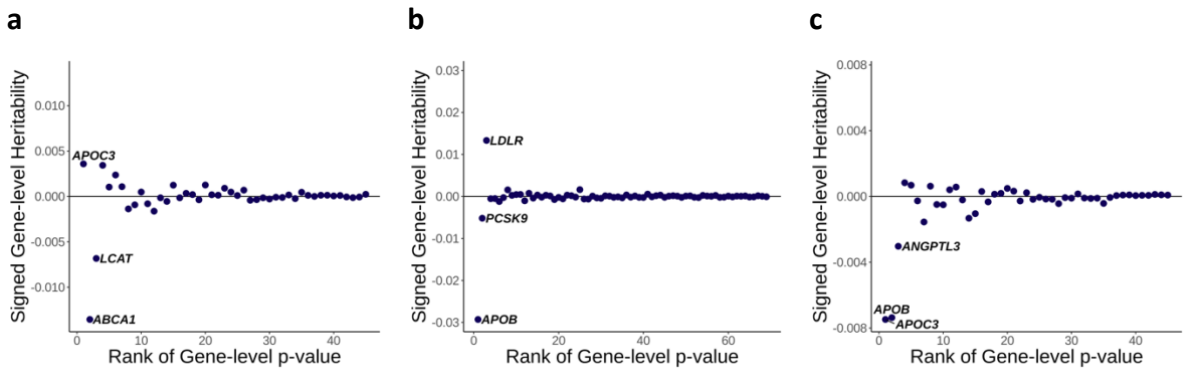
RareEffect encompasses three steps. In step 1, we fit a null GLMM using AI-REML approach, and obtain residuals for the subsequent steps. In step 2, we divide variants by gene and its functional annotation (LoF, missense, and synonymous). We first estimate the variance

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

component of each group and adjust them using MoM approach. In step 3, we calculate the variant-level effect size using BLUP estimates. For binary traits, Firth bias correction is additionally applied to adjust the case-control imbalance. Through RareEffect, we provide region-level and variant-level effect sizes for enhanced interpretability and improved risk prediction performance.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

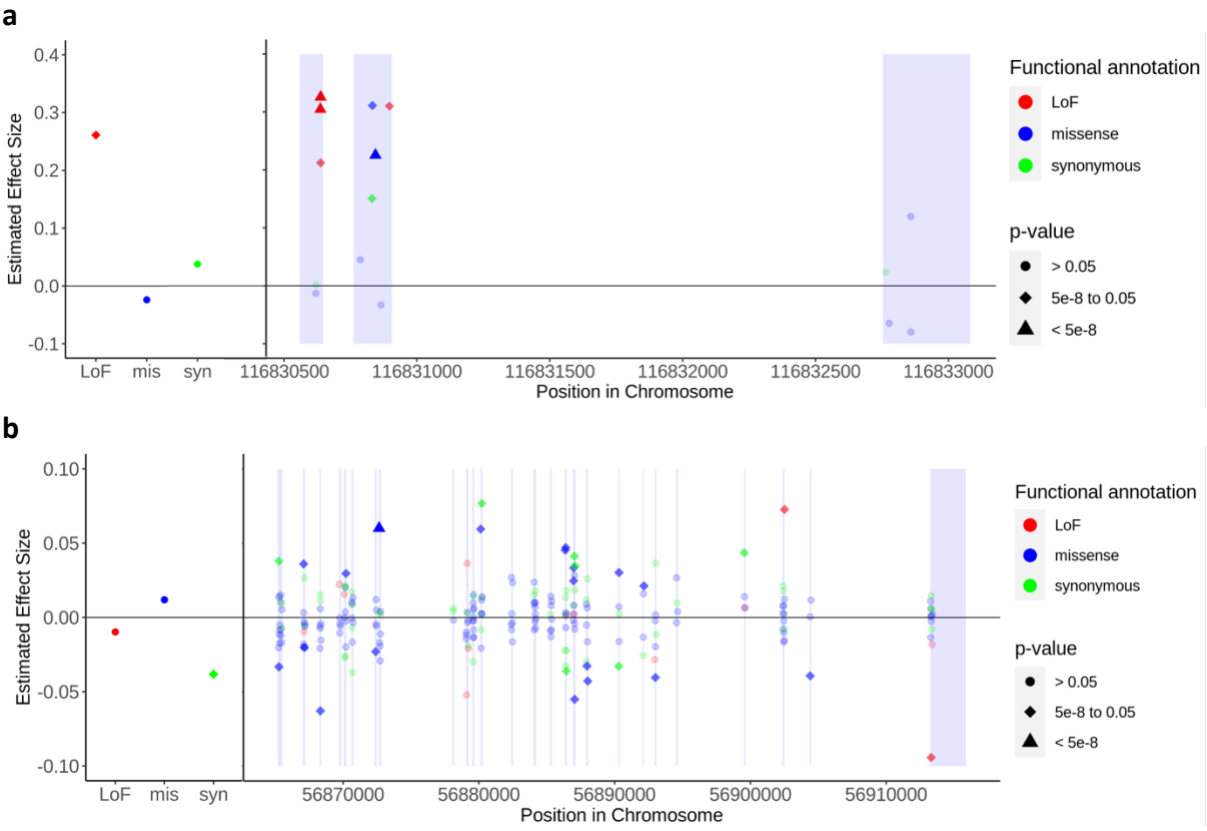
Fig. 2. Estimated signed heritability for lipid phenotypes using 392,748 White British samples in UK Biobank whole exome sequencing data



Signed gene-level heritability from RareEffect for (a) HDL cholesterol level, (b) LDL cholesterol level, and (c) triglycerides level. Gene-level p-values were obtained from SAIGE-GENE+, and we included genes with p-values < 2.5×10^{-6}. The x-axis represents the rank order of genes based on their gene-level p-values. Lower ranks correspond to genes with more significant p-values. The y-axis shows the signed gene-level heritability for each gene. Signed heritability indicates the direction (positive or negative) and magnitude of the genetic contribution of the gene to the trait.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Fig. 3. Variant-level effect size on HDL cholesterol for variants in *APOC3* gene (chromosome 11) and *SLC12A3* gene (chromosome 16).

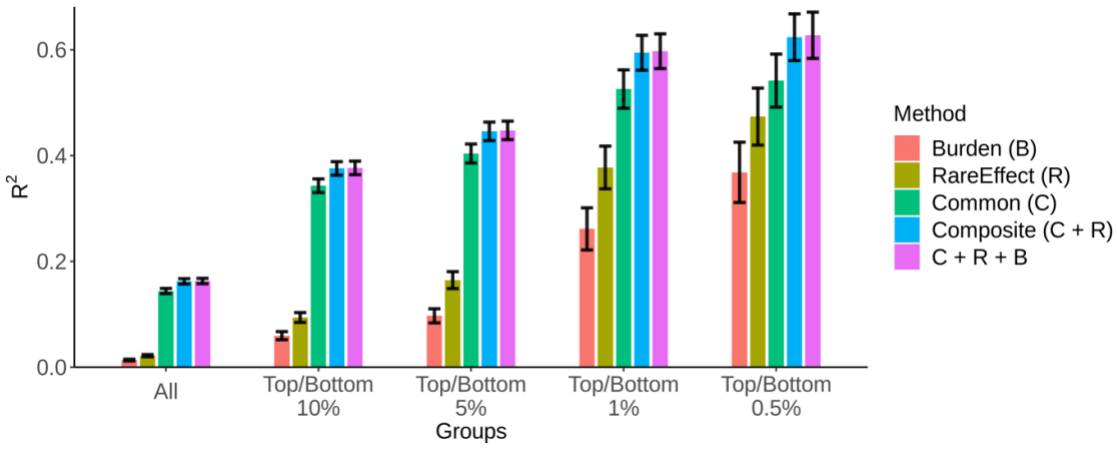


Variant-level effect size on HDL cholesterol level for (a) *APOC3* and (b) *SLC12A3* genes. Gene-level p-values were obtained from SAIGE-GENE+, and we included genes with p-values < 2.5×10^{-6}. The left panel shows the effect size of collapsed ultra-rare variants categorized by their functional annotations: loss-of-function (LoF), missense, and synonymous, respectively. The right panel displays the variant-level effect size of rare variants. Variants are color-coded based on their functional annotation: red for LoF, blue for missense, and green for synonymous. The shapes of the points indicate the significance of the variants: circles represent p-values > 0.05, diamonds represent p-values between 5×10^{-8} and 0.05, and triangles represent p-values < 5×10^{-8}. Single-variant p-values were obtained from SAIGE, while the p-values of collapsed variants were derived from linear regression by regressing $\tilde{y}$ on each collapsed variant. The exon region is shaded.

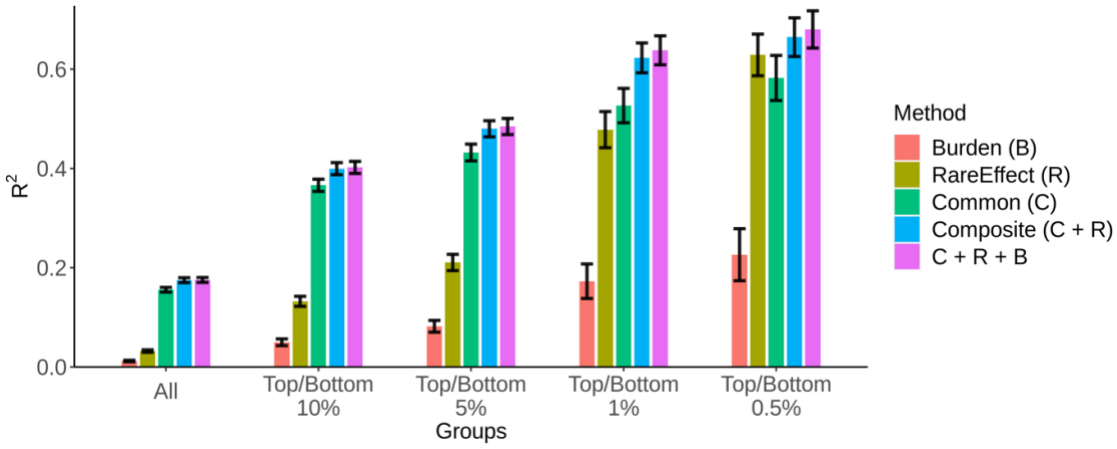
medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Fig. 4. Comparison of performance of risk prediction models for lipid phenotypes using 314,198 White British samples (80% of the whole White British samples) in UK Biobank whole exome sequencing data.

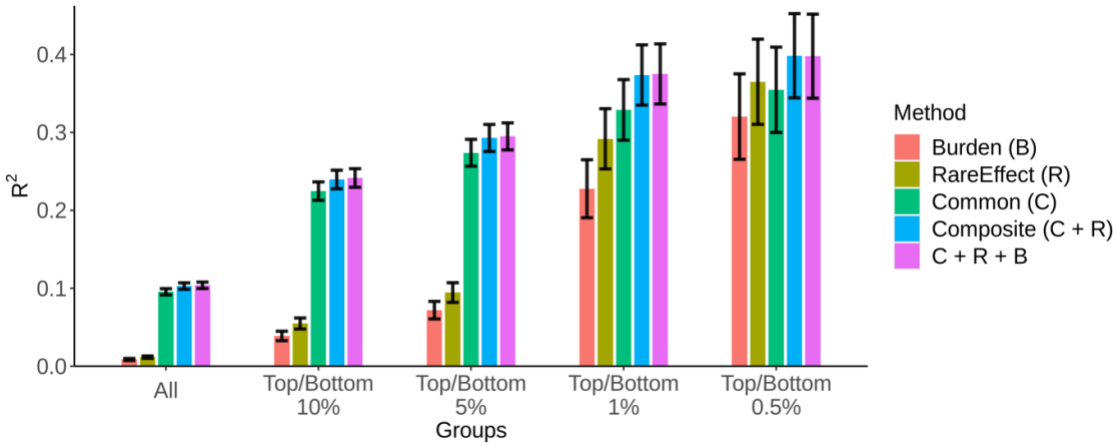
a



b



c



medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Coefficient of determination (R^2) of the risk prediction models for (a) HDL cholesterol level, (b) LDL cholesterol level, and (c) Triglycerides level. We evaluated the R^2 of five models using: (1) PRS_{burden} only, (2) PRS_{RE} only, (3) PRS_{common} only, (4) composite score, (5) $PRS_{common} + PRS_{RE} + PRS_{burden}$ by five subgroups: (1) all samples, (2) samples with top/bottom 10% PRS, (3) samples with top/bottom 5% PRS, (4) samples with top/bottom 1% PRS, and (5) samples with top/bottom 0.5% PRS. The black vertical lines represent the 95% confidence interval of the R^2 estimates.

medRxiv preprint doi: https://doi.org/10.1101/2024.06.23.24309366; this version posted June 24, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC 4.0 International license .

Supplementary Figures

1. Estimated signed heritability for 10 tested traits
2. Predictive performance for 10 tested traits
3. Relationship between common PRS (PRS_{common}) and RareEffect PRS (PRS_{RE})
4. Enrichment of high-risk individuals in terms of RareEffect PRS (PRS_{RE}), common variant PRS (PRS_{common}) and composite score in phenotype outliers
5. Computation time of RareEffect, linear regression, and ridge regression
6. Memory Usage of RareEffect by number of variants
7. Computation time of the fast implementation Firth bias correction and the normal Firth correction
8. Relationship between RareEffect PRS (PRS_{RE}) and phenotype values
9. Comparison of estimated variance components between RareEffect and MoM estimator
10. Comparison of estimated effect size between computing the hat matrix at every iteration and only once in Firth bias correction

Supplementary Tables

1. Predictive performance for 10 tested phenotypes in UK Biobank
2. Pearson correlation between PRS_{common} and PRS_{RE}
3. Predictive performance (RMSE) for simulated continuous data by scenario
4. Predictive performance (RMSE) for simulated binary data by scenario